

Docket No.: 67161-141

**PATENT**

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Application of : Customer Number: 20277

Ryuji MANO : Confirmation Number:

Serial No.: : Group Art Unit:

Filed: February 12, 2004 : Examiner:

For: SPEECH RECOGNITION APPARATUS CAPABLE OF IMPROVING RECOGNITION RATE  
REGARDLESS OF AVERAGE DURATION OF PHONEMES

**CLAIM OF PRIORITY AND  
TRANSMITTAL OF CERTIFIED PRIORITY DOCUMENT**

Mail Stop CPD  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

Sir:

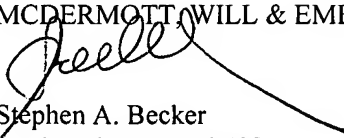
In accordance with the provisions of 35 U.S.C. 119, Applicant hereby claim the priority of:

**Japanese Patent Application No. JP 2003-277661, filed on July 22, 2003.**

cited in the Declaration of the present application. A certified copy is submitted herewith.

Respectfully submitted,

MCDERMOTT, WILL & EMERY

  
Stephen A. Becker  
Registration No. 26,527

600 13<sup>th</sup> Street, N.W.  
Washington, DC 20005-3096  
(202) 756-8000 SAB:gav  
Facsimile: (202) 756-8087  
**Date: February 12, 2004**

67161-141  
Ryuji MANO  
February 12, 2004

日 本 国 特 許 庁 *McDermott, Will & Emery*  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日                      2 0 0 3 年    7 月 2 2 日  
Date of Application:

出 願 番 号                      特 願 2 0 0 3 - 2 7 7 6 6 1  
Application Number:

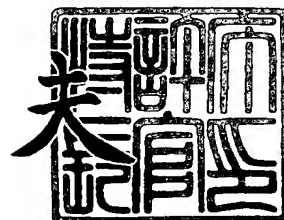
[ST. 10/C] :                      [ J P 2 0 0 3 - 2 7 7 6 6 1 ]

出      願      人                      株式会社ルネサステクノロジ  
Applicant(s):

2 0 0 3 年    8 月 2 6 日

特許庁長官  
Commissioner,  
Japan Patent Office

今 井 康



出証番号    出証特 2 0 0 3 - 3 0 6 9 6 6 1

【書類名】 特許願  
【整理番号】 544000JP01  
【提出日】 平成15年 7月22日  
【あて先】 特許庁長官殿  
【国際特許分類】 G10L 3/00  
G10L 15/02

【発明者】  
【住所又は居所】 東京都千代田区丸の内二丁目 4 番 1 号 株式会社ルネサステクノ  
ロジ内  
【氏名】 眞野 竜二

【特許出願人】  
【識別番号】 503121103  
【氏名又は名称】 株式会社ルネサステクノロジ

【代理人】  
【識別番号】 100064746  
【弁理士】  
【氏名又は名称】 深見 久郎

【選任した代理人】  
【識別番号】 100085132  
【弁理士】  
【氏名又は名称】 森田 俊雄

【選任した代理人】  
【識別番号】 100083703  
【弁理士】  
【氏名又は名称】 仲村 義平

【選任した代理人】  
【識別番号】 100096781  
【弁理士】  
【氏名又は名称】 堀井 豊

【選任した代理人】  
【識別番号】 100098316  
【弁理士】  
【氏名又は名称】 野田 久登

【選任した代理人】  
【識別番号】 100109162  
【弁理士】  
【氏名又は名称】 酒井 將行

【手数料の表示】  
【予納台帳番号】 008693  
【納付金額】 21,000円

【提出物件の目録】  
【物件名】 特許請求の範囲 1  
【物件名】 明細書 1  
【物件名】 図面 1  
【物件名】 要約書 1

**【書類名】 特許請求の範囲****【請求項 1】**

入力音声信号に対して、各々が所定時間長を有する時間窓に対応する複数のフレームを、少なくとも異なる時間幅でスライドさせることにより、特徴パラメータの抽出を行う特徴抽出手段と、

前記入力音声の音韻パターンにそれぞれ対応づけて標準パターンデータを格納するための記憶手段と、

前記特徴抽出手段で抽出された前記特徴パラメータと前記標準パターンデータを照合することで、対応する音韻を認識し、認識結果を出力するための認識手段とを備える、音声認識装置。

**【請求項 2】**

前記特徴抽出手段は、前記入力音声信号の語頭から語尾に渡って、前記フレームをスライドさせる時間幅を順次増加させ、

前記記憶手段は、前記特徴抽出手段が前記フレームをスライドさせる時間幅のパターンに対応する前記標準パターンデータを予め格納する、請求項 1 記載の音声認識装置。

**【請求項 3】**

前記特徴抽出手段は、

第 1 の固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する第 1 の固定フレーム間隔抽出処理手段と、

前記第 1 の固定時間幅よりも短い第 2 の固定時間幅で前記時間窓をスライドさせつつ、前記特徴パラメータを抽出する第 2 の固定フレーム間隔抽出処理手段とを含み、

前記標準パターンデータは、前記第 1 の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 1 のパターンに対応する第 1 の標準パターンデータと、前記第 2 の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 2 のパターンに対応する第 2 の標準パターンデータとを含む、請求項 1 記載の音声認識装置。

**【請求項 4】**

前記特徴抽出手段は、

固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する固定フレーム間隔抽出処理手段と、

前記入力音声信号の語頭から語尾に渡って、前記フレームをスライドさせる時間幅を順次増加させつつ、前記特徴パラメータを抽出する可変フレーム間隔抽出処理手段とを含み、

前記標準パターンデータは、前記固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 1 のパターンに対応する第 1 の標準パターンデータと、前記可変フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 2 のパターンに対応する第 2 の標準パターンデータとを含む、請求項 1 記載の音声認識装置。

**【請求項 5】**

前記特徴抽出手段は、

第 1 の固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する第 1 の固定フレーム間隔抽出処理手段と、

前記第 1 の固定時間幅よりも短い第 2 の固定時間幅で前記時間窓をスライドさせつつ、前記特徴パラメータを抽出する第 2 の固定フレーム間隔抽出処理手段とを含み、

前記標準パターンデータは、前記第 1 の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 1 のパターンに対応する第 1 の標準パターンデータと、前記第 2 の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 2 のパターンに対応する第 2 の標準パターンデータとを含み、

前記入力音声信号と前記特徴抽出手段との間に設けられ、前記第 1 の固定フレーム間隔抽出処理手段から抽出された前記特徴パラメータに基づく、前記認識手段による照合結果に応じて、前記入力音声信号を前記第 1 の固定フレーム間隔抽出処理手段から前記第 2 の固定フレーム間隔抽出処理手段に切替えて与える入力選択手段をさらに備える、請求項 1

記載の音声認識装置。

【請求項 6】

前記第 1 の標準パターンデータは時刻と関連付けられており、

前記第 1 の標準パターンデータに基づいて、前記第 2 の標準パターンデータを補間により生成する補間処理手段をさらに備える、請求項 5 記載の音声認識装置。

【請求項 7】

前記第 1 の標準パターンデータおよび第 2 の標準パターンデータは時刻と関連付けられており、

前記第 2 の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる各時刻点は、前記第 1 の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時刻点のいずれかに対応する、請求項 6 記載の音声認識装置。

【請求項 8】

前記特徴抽出手段は、

固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する固定フレーム間隔抽出処理手段と、

前記入力音声信号の語頭から語尾に渡って、前記フレームをスライドさせる時間幅を順次増加させつつ、前記特徴パラメータを抽出する可変フレーム間隔抽出処理手段とを含み、

前記標準パターンデータは、前記固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 1 のパターンに対応する第 1 の標準パターンデータと、前記可変フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第 2 のパターンに対応する第 2 の標準パターンデータとを含み、

前記入力音声信号と前記特徴抽出手段との間に設けられ、前記固定フレーム間隔抽出処理手段から抽出された前記特徴パラメータに基づく、前記認識手段による照合結果に応じて、前記入力音声信号を前記固定フレーム間隔抽出処理手段から前記可変フレーム間隔抽出処理手段に切替えて与える入力選択手段をさらに備える、請求項 1 記載の音声認識装置。

【請求項 9】

前記第 1 の標準パターンデータは時刻と関連付けられており、

前記第 1 の標準パターンデータに基づいて、前記第 2 の標準パターンデータを補間により生成する補間処理手段をさらに備える、請求項 8 記載の音声認識装置。

**【書類名】明細書****【発明の名称】音声認識装置****【技術分野】****【0001】**

本発明は、音韻単位の認識に基づく音声認識装置の構成に関するものである。

**【背景技術】****【0002】**

従来、音声認識装置における音声認識は、そのほとんどが音声の特徴量の時系列に変換し、その時系列をあらかじめもっている標準パターンの時系列と比較して認識を行うことにより実行されている。

**【0003】**

たとえば、特開2001-356790号公報では、人の音声を機械に認識させる音声認識装置において、特徴量抽出部が、分析対象音声から、所定の周期毎に設定された一定長の複数の時間窓から音声特徴量を抽出する技術が開示されている。この技術では、音声の周波数に関する周波数軸系特徴パラメータと、音声の振幅に関するパワー系特徴パラメータとを、それぞれ異なる周期で抽出する。

**【0004】**

また、特開平5-303391号公報では、特徴パラメータを計算するための単位時間（フレーム）を複数個用意する、あるいは各音韻毎に用意し、各フレーム長毎に特徴パラメータ時系列を計算し、そのそれぞれに対して音韻照合を行い、最適なものを選ぶ技術が開示されている。

【特許文献1】特開2001-356790号公報明細書

【特許文献2】特開平5-303391号公報明細書

**【発明の開示】****【発明が解決しようとする課題】****【0005】**

ただし、上述したような、一定長の複数の時間窓を一定時間ごとにずらしながら音声の特徴量の時系列に変換するという方法では、音韻の長さに応じて、抽出される特徴パラメータの数が異なってしまう。その結果、この特徴パラメータの数が、認識率に影響を与えてしまうという問題があった。

**【0006】**

本発明は、上記のような問題を解決するためになされたものであって、その目的は、各音韻の認識率を向上することが可能な特徴パラメータの計算方法を用いた音声認識装置を提供することである。

**【課題を解決するための手段】****【0007】**

このような目的を達成するために、本発明の音声認識装置は、入力音声信号に対して、各々が所定時間長を有する時間窓に対応する複数のフレームを、少なくとも異なる時間幅でスライドさせることにより、特徴パラメータの抽出を行う特徴抽出手段と、入力音声の音韻パターンにそれぞれ対応づけて標準パターンデータを格納するための記憶手段と、特徴抽出手段で抽出された特徴パラメータと標準パターンデータを照合することで、対応する音韻を認識し、認識結果を出力するための認識手段とを備える。

**【発明の効果】****【0008】**

本発明の音声認識装置では、音韻の平均継続時間長が長い場合でも、あるいは、短い場合でも、処理負荷を低減しつつ、各音韻の認識率を向上することが可能である。

**【発明を実施するための最良の形態】****【0009】**

以下、図面を参照して本発明の実施の形態について説明する。

**【0010】**

(本発明の構成の説明の前提)

以下では、まず、本発明の音声認識装置の構成を説明する前提として、一般的な音声認識装置 10 の構成および動作について、予め説明しておく。

【0011】

図 1 は、このような音声認識装置 10 の構成を説明するための機能ブロック図である。

【0012】

図 1 を参照して、特徴検出部 102 は、入力として与えられた入力音声 101 に対して、LPC ケプストラム係数（数十ミリ秒の音声切り出し単位であるフレームごとの対数パワースペクトル包絡のフーリエ変換）等の特徴パラメータを計算する。すなわち、特徴検出部 102 は、特徴量を計算する場合、通常数ミリ秒から数十ミリ秒を単位時間（フレーム）とし、1 フレームの時間内では特徴量すなわち音声の波の構造は定常状態にあると近似して、特徴パラメータを計算する。そして、フレームをある時間だけずらして（これをフレームシフトという）、ふたたび特徴パラメータを計算する。これを繰り返すことによって特徴パラメータの時系列が得られる。

【0013】

認識部 103 は、このようにして得られた特徴パラメータの時系列を、記憶装置に格納された単語辞書データベース（単語辞書 DB）104 内の標準パターンと比較し、類似度を計算することによって、認識結果 105 を出力する。

【0014】

図 2 は、図 1 に示した特徴検出部 102 におけるフレームシフトを説明するための概念図である。

【0015】

図 2 に示すように、音声認識装置 10 の特徴検出部 102 では、フレームシフトの時間幅 D201 は、一定である。このため、音韻の継続時間の長い単語と、短い単語で、特徴パラメータの数が異なることになる。したがって、音韻の長い単語は、認識率が良く、音韻の短い単語は、認識率が音韻の長い単語に比べて悪いという傾向が生じる。

【0016】

本発明においては、以下に説明するように、フレームシフトの時間幅を可変にして、特徴パラメータを計算することで、音韻の解析を左右するとされる箇所に重点をおいて、音韻の長い単語と、音韻の短い単語とで、特徴パラメータの生成数を同数とするように特徴量の抽出を行う。

【0017】

〔実施の形態 1〕

以下、本発明の実施の形態 1 の音声認識装置 100 の構成および動作について説明する。

【0018】

図 3 は、音声認識装置 100 の構成を説明するための機能ブロック図である。

【0019】

図 3 に示した音声認識装置 100 の構成は、基本的には、図 1 に示した音声認識装置 10 の構成と同様である。

【0020】

ただし、以下に説明するとおり、まず、発話者の音声デジタル化されたものである入力音声 301 を受ける特徴抽出部 302 において、特徴パラメータ計算部 3021 は、フレームシフトの間隔を音韻の語頭のフレーム間隔ほど密にし、語尾に向けて順次フレーム間隔を粗くすることで、特徴パラメータを計算する。さらに、このようにして計算された特徴パラメータの時系列を受けて認識処理部 303 が認識処理を行うにあたり、参照する単語辞書データベース 304 が、このような可変なフレーム間隔に対応するように、以下に説明するとおり、所定の規則で変化するフレーム間隔に応じた標準パターンを予め格納する構成となっている。認識処理部 303 は、このような単語辞書データベース 304 を参照して、特徴パラメータの時系列との照合を行って解析することにより、認識結果 30

5 を出力する。

【0021】

以下、音声認識装置 100 の動作についてさらに詳しく説明する。

【0022】

音韻認識をおこなう場合、それぞれの音韻の平均継続時間長が重要である。音韻の特徴は、大きく語頭、話中、語尾の 3 つに分けられる。発音記号の /t/ や /r/ で表される子音は、語頭・話中・語尾の平均継続時間長は 15 ミリ秒程度しかないのに対し、母音の方はそれぞれ 100 ミリ秒を越える平均時間長をもつ。このように継続時間長に大きなばらつきがある様々な音素を認識するにあたり、単語の先頭データの重要度が特に高い。このため、本発明では、フレームシフトの時間幅を、以下に説明する所定の規則に基づいて変化させる。

【0023】

図 4 は、音声認識装置 100 の特徴パラメータ計算部 3021 のフレームシフト動作を説明するための概念図である。

【0024】

例えば、図 4 においては、サンプリング周波数 20 キロヘルツで 16 ビットで量子化された入力音声 301 から、特徴パラメータ計算部 3021 において、特徴パラメータを計算するものとする。

【0025】

特徴パラメータ計算部 3021 は、時間窓である固定のフレーム長  $L$  を、入力音声の先頭から、終わりまで、順次長くなっていく時間幅  $D301 \sim D30n$  (例:  $D301 < D302 < D303 < \dots < D30n$ ,  $n$ : 自然数) でシフトし、それぞれ特徴パラメータ時系列  $S1 \sim Sn$  を生成する。

【0026】

ここで、特に限定されないが、たとえば、時間幅  $D301 \sim D30n$  を順次長くしていくにあたっては、たとえば、先頭のフレームから次のフレームまでの時間間隔  $D301$  を基準として、所定の割合で等比級数的に以後の時間間隔  $D302 \sim D30n$  を順次長くしていくことや、所定の間隔で等差級数的に以後の時間間隔  $D302 \sim D30n$  を順次長くしていくことが可能である。もちろん、より一般的に、時間に対して単調に増加する関数にしたがって、時間間隔  $D302 \sim D30n$  を順次長くしていくことも可能である。

【0027】

まず、この入力音声 301 の先頭からフレーム長  $L$  分のデータに注目し、この中のデータを定常状態にあるとみなして特徴パラメータを計算する。たとえば、12 次の線形予測係数 LPC (Linear Predictive Coding) から 16 次の LPC ケプストラム係数を計算して 16 次元の特徴ベクトルとする。次にフレームを時間幅  $D30i$  ( $i = 1 \sim n$ ) でシフトさせ、同様に特徴ベクトルを計算する。この換作を音声入力 301 の終わりまでくりかえすことによって、固定のフレーム長  $L$  を用いて計算した特徴パラメータ時系列  $Sn$  が得られる。

【0028】

特徴パラメータ計算部 3021 からの特徴パラメータの出力に対しては、認識処理部 303 において、フレームごとに、単語辞書データベース 304 とのパラメータ比較が行われる。全フレームの比較が行われ、単語辞書データベース 304 に登録されたモデルのうちで、しきい値を満たす最も適するものが、認識結果 305 として出力される。

【0029】

このとき、単語辞書データベース 304 へ格納するデータとしては、予め個々の音韻モデルに関して、フレーム長  $L$  において時間幅  $D301 \sim D30n$  でフレームシフトして計算した特徴パラメータを用いて、標準パターンを作成しておく。このような標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用い、計算した特徴パラメータ時系列を、個別の隠れマルコフモデル (HMM: Hidden Markov Model) P01 を用意してトレーニングすることによって作成される。こうして、得られた音韻数  $M$  ( $M$



: 所定の自然数) の隠れマルコフモデルにより、単語辞書データベース 104 が構成される。

#### 【0030】

認識処理部 304 では、音韻認識を行うにあたり、すべての音韻の存在位置・存在確率を調べ、存在位置が重なっているものに関しては存在確率の大きなもののみを残す。こうして得られた音韻列を認識結果 105 として出力するものとする。

#### 【0031】

以上のような構成を有する音声認識装置 100 により、フレームシフトの時間幅を固定した場合の音韻認識率と比較して、音韻の語頭に対する特徴パラメータの重み付けを大きくすることで、認識率を向上させることが可能となる。

#### 【0032】

##### [実施の形態 2]

図 5 は、実施の形態 2 の音声認識装置 200 の構成を説明するための機能ブロック図である。

#### 【0033】

なお、以下では、図 2 に示したように、時間窓であるフレーム間の間隔を固定して、特徴パラメータを抽出する処理手順を「固定フレーム間隔抽出処理」と呼ぶこととする。

#### 【0034】

図 5 に示した音声認識装置 200 は、デジタル化された入力音声 401 に対して、第 1 の時間間隔での固定フレーム間隔抽出処理を行う第 1 の特徴パラメータ計算部を有する第 1 の特徴抽出部 402 と、第 2 の時間間隔での固定フレーム間隔抽出処理を行う第 2 の特徴パラメータ計算部を有する第 2 の特徴抽出部 403 とを備える。

#### 【0035】

第 1 の特徴抽出部 402 および第 2 の特徴抽出部 403 にて、それぞれ第 1 の特徴パラメータ時系列  $S_{01} \sim S_{0n}$  および第 2 の特徴パラメータ時系列  $S_{11} \sim S_{1n}$  を計算する。

#### 【0036】

また、音声認識装置 200 は、予め第 1 の時間間隔での固定フレーム間隔抽出処理に対応した音韻モデルが登録された第 1 の単語辞書データベース 4022 と、予め第 2 の時間間隔での固定フレーム間隔抽出処理に対応した第 2 の単語辞書データベース 4032 と、第 1 の特徴抽出部 402 で計算された特徴パラメータのそれぞれを第 1 の単語辞書データベース 4022 内のデータと比較して音韻を認識するための第 1 の認識処理部 4021 と、第 2 の特徴抽出部 403 で計算された特徴パラメータのそれぞれを第 2 の単語辞書データベース 4032 内のデータと比較して音韻を認識するための第 2 の認識処理部 4031 と、さらに、第 1 および第 2 の認識処理部 4021, 4031 の認識結果を、その適合度に応じて選択し、認識結果 405 を得るための結果選択部 404 とを備える。

#### 【0037】

以下、音声認識装置 200 の動作について、さらに詳しく説明する。

#### 【0038】

まず、入力音声 401 の先頭からフレーム長  $L$  分のデータに注目し、この中のデータを定常状態にあるとみなして、第 1 の特徴抽出部 402 および第 2 の特徴抽出部 403 にて、特徴パラメータを計算する。

#### 【0039】

音声認識装置 200 では、第 1 の特徴抽出部 402 において、たとえば、12 次の線形予測係数  $LPC$  から 16 次の  $LPC$  ケプストラム係数を計算して 16 次元の特徴ベクトルとする。同様に、第 2 の特徴抽出部 403 においても、12 次の線形予測係数  $LPC$  から 16 次の  $LPC$  ケプストラム係数を計算して 16 次元の特徴ベクトルとする。

#### 【0040】

その結果、第 1 の特徴抽出部 402 および第 2 の特徴抽出部 403 のそれぞれにおいて、第 1 の特徴パラメータ  $S_{01}$ 、第 2 の特徴パラメータ  $S_{11}$  が得られる。この操作以降

、入力音声 401 の信号の終わりまで、第 1 の特徴抽出部 402 では、固定時間幅 D201 でフレームシフトを繰り返し計算した第 1 の特徴パラメータ S0n を出力し、第 2 の特徴抽出部 403 では、固定時間幅 D2011 (<D201) でフレームシフトを繰り返し計算した第 2 の特徴パラメータ S1n を出力する。

#### 【0041】

一方、あらかじめ個々の音韻モデルに関して、フレーム長 L から計算した特徴パラメータを用いて、第 1 の標準パターンを作成しておく。この第 1 の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用いて計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅を D201 にして、作成したものである）を、個別の隠れマルコフモデル（HMM）P01 を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数 M の隠れマルコフモデルにより、第 1 の単語辞書データベース 4022 が構成される。

#### 【0042】

また、第 2 の標準パターンも同様に、あらかじめフレーム長 L から計算した特徴パラメータを用いて、作成しておく。この第 2 の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用い、計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅を D2011 にして、作成したものである）を、個別の隠れマルコフモデル（HMM）P11 を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数 M の隠れマルコフモデルにより、第 2 の単語辞書データベース 4032 が構成される。

#### 【0043】

第 1 の認識処理部 4021 においては、入力音声の先頭のフレームから順に各音韻毎に特徴パラメータ時系列 S01 は標準パターン P01 を用い、特徴パラメータ時系列 S02 には標準パターン P02 を用いて照合を行ない、以下同様にして、特徴パラメータ時系列 S0n には標準パターン P0n を用いて音韻照合を行い、存在位置および存在確率の重なるものを出力する。

#### 【0044】

同様に、第 2 の認識処理部 4031 においては、入力音声の先頭のフレームから順に各音韻毎に特徴パラメータ時系列 S11 は標準パターン P11 を用い、特徴パラメータ時系列 S12 には標準パターン P12 を用いて照合を行ない、以下同様にして、特徴パラメータ時系列 S1n には標準パターン P1n を用いて音韻照合を行い、存在位置および存在確率の重なるものを出力する。

#### 【0045】

結果選択部 404 では、第 1 の認識処理部 4021 および第 2 の認識処理部 4031 から出力されたすべての音韻の存在位置・存在確率を調べ、存在位置が重なっているものに関しては存在確率の大きなもののみを残す。結果選択部 404 は、こうして得られた音韻列を認識結果 405 として出力する。

#### 【0046】

以上説明したような音声認識装置 200 の構成により、フレーム間の時間間隔を固定した場合の音韻認識率と比較して、異なったフレーム間の時間間隔で抽出された特徴パラメータを用いて、より存在確率の高い方が選択されるので、認識率を向上することができる。

#### 【0047】

##### 〔実施の形態 3〕

以下では、図 4 で説明したように、時間窓であるフレーム間の間隔を順次長くしながら、特徴パラメータを抽出する処理手順を「可変フレーム間隔抽出処理」と呼ぶこととする。

#### 【0048】

実施の形態 2 では、第 1 の特徴抽出部 402 と第 2 の特徴抽出部 403 との双方が、固定フレーム間隔抽出処理を行なうものとした。

## 【0049】

これに対して、本発明の実施の形態3の音声認識装置の基本的な構成は、実施の形態2の音声認識装置200の構成と同様である。

## 【0050】

ただし、実施の形態3の音声認識装置では、第2の特徴抽出部403は、可変フレーム間隔抽出処理を行なうものとする。

## 【0051】

すなわち、第2の特徴抽出部403は、図4で説明したようにフレームシフトの時間幅  $D30i$  ( $i$ : 自然数、 $D301 < D302 < D303 < \dots$ ) を順次長くしながら可変にし、特徴パラメータをそれぞれにおいて計算する。

## 【0052】

また、第2の単語辞書データベース4032には、フレームシフトの時間幅を  $D30i$  ( $i$ : 自然数、 $D301 < D302 < D303 < \dots$ ) にして計算した特徴パラメータを用いて、標準パターンを作成しておくものとする。

## 【0053】

実施の形態3の音声認識装置のその他の構成は、実施の形態2の音声認識装置200の構成と同様であるので、その説明は繰り返さない。

## 【0054】

このような実施の形態3の音声認識装置の構成により、音声認識装置200の奏する効果に加え、音韻の平均継続時間長が長い場合は、固定フレーム間隔抽出処理で有効に対処することが可能であり、一方、音韻の平均継続時間長が短い場合は、可変フレーム間隔抽出処理で有効に対処することが可能であるので、処理負荷を低減できる。

## 【0055】

## [実施の形態4]

図6は、実施の形態4の音声認識装置300の構成を説明するための機能ブロック図である。

## 【0056】

図6に示した音声認識装置300は、デジタル化された入力音声501に対して、第1の時間間隔での固定フレーム間隔抽出処理を行う第1の特徴パラメータ計算部を有する第1の特徴抽出部502と、第2の時間間隔での固定フレーム間隔抽出処理を行う第2の特徴パラメータ計算部を有する第2の特徴抽出部503とを備える。

## 【0057】

さらに、音声認識装置300は、後に説明する制御信号51を入力とするインバータ511と、制御信号51およびインバータ511の出力信号50に応じて、入力音声501を、第1の特徴抽出部502または第2の特徴抽出部503に選択的に与えるための入力選択部510を備える。

## 【0058】

入力選択部510は、入力音声501および制御信号51を入力に受け、出力を第1の特徴抽出部502に与えるAND回路512と、入力音声501およびインバータ511の出力信号50を入力に受け、出力を第2の特徴抽出部503に与えるAND回路513とを備える。

## 【0059】

第1の特徴抽出部502および第2の特徴抽出部503にて、それぞれ第1の特徴パラメータ時系列  $S01 \sim S0n$  および第2の特徴パラメータ時系列  $S11 \sim S1n$  を計算する。

## 【0060】

また、音声認識装置300は、予め第1の時間間隔での固定フレーム間隔抽出処理に対応した音韻モデルが登録された第1の単語辞書データベース5022と、予め第2の時間間隔での固定フレーム間隔抽出処理に対応した第2の単語辞書データベース5032と、第1の特徴抽出部502で計算された特徴パラメータのそれぞれを第1の単語辞書データ

ベース5022内のデータと比較して音韻を認識するための第1の認識処理部5021と、第2の特徴抽出部503で計算された特徴パラメータのそれぞれを第2の単語辞書データベース5032内のデータと比較して音韻を認識するための第2の認識処理部5031と、さらに、第1および第2の認識処理部5021、5031の認識結果を、以下に説明する手順にしたがって選択し、認識結果505を得るための結果選択部504とを備える。

#### 【0061】

結果選択部504は、第1の認識処理部5021の出力および制御信号51を入力に受け、認識結果505を出力するAND回路514と、第2の認識処理部5031の出力および出力信号50を入力に受け、認識結果505を出力するAND回路515とを備える。

#### 【0062】

以下、音声認識装置300の動作について説明する。

#### 【0063】

まず、入力音声501の先頭からフレーム長L分のデータに注目し、この中のデータを定常状態にあるとみなし、制御信号51に応じて第1の特徴抽出部502、あるいは第2の特徴抽出部503において特徴パラメータを計算する。

#### 【0064】

ここで、制御信号51は、第1の認識処理部5021における認識処理で、認識結果を得るために設定したしきい値を満たす場合は、第1の特徴抽出部502に音声を入力し、第1の認識処理部5021ではしきい値を満たさない場合には、第2の特徴抽出部503に音声を入力するように変化するものとする。

#### 【0065】

例えば、入力音声501が、登録単語のいくつかと、語頭は同じであるが、語尾になると異なるような場合、第1の特徴抽出部502および第1の認識処理部5021からなる第1の処理系で、語頭から語尾にかけてフレームごとに認識処理を行うにつれて、次第にしきい値を満たさなくなっていくことが起り得る。

#### 【0066】

このとき、第1の認識処理部5021は、制御フラグを制御信号51として返し、そのフラグによって、第2の特徴抽出部503および第2の認識処理部5031からなる第2の処理系に認識処理を切り替え、シフト時間幅を変化させて認識処理を行うものとする。

#### 【0067】

実施の形態4では、上述した第2の処理系でフレームシフトの時間幅が、第1の処理系でのフレームシフトの時間幅よりも短いものであることとして、以下説明する。

#### 【0068】

実施の形態4において、第1の特徴抽出部502および第2の特徴抽出部503においては、12次の線形予測係数LPCから16次のLPCケプストラム係数を計算して16次元の特徴ベクトルとするものとする。

#### 【0069】

その結果、第1の特徴抽出部502および第2の特徴抽出部503のそれぞれにおいて、第1の特徴パラメータS01、第2の特徴パラメータS11が得られる。この操作以降、入力信号の終わりまで、第1の特徴抽出部502では、一定値に固定した時間幅D201でフレームシフトを繰り返し、計算した第1の特徴パラメータS0nを出力し、第2の特徴抽出部503では、固定時間幅D2011 (< D201) でフレームシフトを繰り返し計算した第2の特徴パラメータS1nを出力する。

#### 【0070】

また、第1および第2の単語辞書データベース5022および5032には、実施の形態2と同様にして、フレームシフトの時間幅をD201にして作成した特徴パラメータ時系列およびフレームシフトの時間幅をD2011にして作成した特徴パラメータ時系列のそれぞれに対応した、各音韻モデルに対する隠れマルコフモデルによる第1および第2の

標準パターンが格納されているものとする。

#### 【0071】

第1の認識処理部5021においては、入力音声の先頭のフレームから順にフレーム毎に、特徴パラメータ時系列S01には標準パターンP01を用い、特徴パラメータ時系列S02には標準パターンP02を用いる。以下同様に、第1の認識処理部5021は、特徴パラメータ時系列S0xには標準パターンP0x (x:自然数)を用い、存在位置、存在確率の重なり、設定するしきい値を満たすものを出力する。この処理を繰り返す中で、設定したしきい値を満たさなければ、第1の認識処理部5021は、切り替え信号を生成して制御信号51を反転させ、第2の特徴抽出部503の出力を用いて、第2の認識処理部5031において音韻照合を行なうように処理を切替える。すなわち、以後、第2の認識処理部5031は、同様に、フレーム毎に特徴パラメータ時系列S1 (x+1)には標準パターンP1 (x+1)を用い、特徴パラメータ時系列S1 (x+2)には標準パターンP1 (x+2)を用い、以下同様に、特徴パラメータ時系列S1nには標準パターンP1nを用いて音韻照合を行い、存在位置、存在確率の重なるものを出力する。

#### 【0072】

そして、結果選択部504は、第1または第2の処理系の結果から得られた音韻列を最終的な認識結果505として出力する。

#### 【0073】

以上説明したような実施の形態4の音声認識装置300の構成により、フレームの時間幅を単一に固定した場合の音韻認識率と比較して、認識率を向上させることが可能である。

#### 【0074】

なお、もう一つの効果として、例えば、図示しないもう一つ別の処理系があり、その処理系は特定のものと限定しないが、その図示しない処理系が処理中であるということを示す信号を生成できるものとし、その生成信号を制御信号51として使用することも可能である。その場合、本音声信号処理装置300を含むシステムにおいて、CPU (Central Processing Unit) などの処理負荷を低減できる。

#### 【0075】

##### 〔実施の形態5〕

実施の形態4では、第1の特徴抽出部502と第2の特徴抽出部503との双方が、固定フレーム間隔抽出処理を行なうものとした。

#### 【0076】

これに対して、本発明の実施の形態5の音声認識装置の基本的な構成は、実施の形態4の音声認識装置300の構成と同様である。

#### 【0077】

ただし、実施の形態5の音声認識装置では、第2の特徴抽出部503は、可変フレーム間隔抽出処理を行なうものとする。

#### 【0078】

すなわち、第2の特徴抽出部503は、図4で説明したようにフレームシフトの時間幅D30i (i:自然数、D301<D302<D303<...)を順次長くしながら可変にし、特徴パラメータをそれぞれにおいて計算する。

#### 【0079】

また、第2の単語辞書データベース5032には、フレームシフトの時間幅をD30i (i:自然数、D301<D302<D303<...)にして計算した特徴パラメータを用いて、標準パターンを作成しておくものとする。

#### 【0080】

実施の形態5の音声認識装置のその他の構成は、実施の形態4の音声認識装置300の構成と同様であるので、その説明は繰り返さない。

#### 【0081】

このような実施の形態5の音声認識装置の構成により、音声認識装置300の奏する効

果に加え、音韻の平均継続時間長が長い場合は、固定フレーム間隔抽出処理で有効に対処することが可能であり、一方、音韻の平均継続時間長が短い場合は、可変フレーム間隔抽出処理で有効に対処することが可能であるので、処理負荷を低減できる。

#### 【0082】

##### 【実施の形態6】

図7は、実施の形態6の音声認識装置400の構成を説明するための機能ブロック図である。

#### 【0083】

図7に示した音声認識装置400においては、入力音声601、入力選択部610、制御信号61、インバータ611、第1の特徴抽出部602、第2の特徴抽出部603、第1の認識処理部6021、第2の認識処理部6031、結果選択部604、第1の単語辞書データベース6022および認識結果605は、それぞれ、実施の形態4の音声認識装置300の入力音声501、入力選択部510、制御信号51、インバータ511、第1の特徴抽出部502、第2の特徴抽出部503、第1の認識処理部5021、第2の認識処理部5031、結果選択部504、第1の単語辞書データベース5022および認識結果505に相当する機能を有している。

#### 【0084】

図7に示した音声認識装置400においては、実施の形態4の音声認識装置300の構成とは異なり、第2の単語辞書データ5032の代わりに、データ補間部6032が設けられている。

#### 【0085】

図7に示した音声認識装置400においても、第2の特徴抽出部503および第2の認識処理部5031からなる第2の処理系でのフレームシフトの時間幅D2011が、第1の特徴抽出部502および第1の認識処理部5021からなる第1の処理系でのフレームシフトの時間幅D201よりも短いものであるとする。

#### 【0086】

ここで、音声認識装置400においても、あらかじめ個々の音韻モデルに関して、フレーム長Lから計算した特徴パラメータを用いて、第1の標準パターンを作成しておく。この第1の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用いて計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅をD201にして、作成したものである）を、個別の隠れマルコフモデル（HMM）P01を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数Mの隠れマルコフモデルにより、第1の単語辞書データベース6022が構成される。

#### 【0087】

図8は、このようにして作成された標準パターンが、第1の単語辞書データベース6022に格納される状態を説明するための概念図である。

#### 【0088】

図8に示すとおり、音韻に対応した隠れマルコフモデルに対して、所定の時間における801～80nの第1の標準パターンは、それぞれ時刻t1～tnにおけるパラメータm1～mnとして構成される。

#### 【0089】

音声認識装置400では、第2の処理系でのフレームシフトの時間幅D2011が、第1の処理系でのフレームシフトの時間幅D201よりも短いものであることから、第2の認識処理部5031で使用されるべき第2の標準パターンとして第1の標準パターンを用いようとしたとしても、第1の単語辞書データベース6022には、第2の標準パターンとしては存在しない部分が生じる。

#### 【0090】

そこで、音声認識装置400では、第2の標準パターンを第1の標準パターンに基づいて、データ補間部6032により生成する。

## 【0091】

図9は、データ補間部6032の処理を説明するための概念図である。

## 【0092】

図9に示すように、第1の標準パターンと時間データを用いて中間データを線形補間（任意の高次関数でも可）によって計算することで、全ての時間における第2の標準パターンを作成できる。

## 【0093】

音声認識装置400のその他の動作は、実施の形態4と同様であるので、その説明は繰り返さない。

## 【0094】

以上のような音声認識装置400の構成とすれば、単語辞書データベースとして使用するメモリ等の記憶装置の記憶容量を削減できる。

## 【0095】

## [実施の形態7]

実施の形態6では、第1の特徴抽出部602と第2の特徴抽出部603との双方が、固定フレーム間隔抽出処理を行なうものとした。

## 【0096】

これに対して、本発明の実施の形態7の音声認識装置の基本的な構成は、実施の形態6の音声認識装置400の構成と同様である。

## 【0097】

ただし、実施の形態7の音声認識装置では、第2の特徴抽出部603は、可変フレーム間隔抽出処理を行なうものとする。

## 【0098】

すなわち、第2の特徴抽出部603は、図4で説明したようにフレームシフトの時間幅  $D30i$  ( $i$ : 自然数、 $D301 < D302 < D303 < \dots$ ) を順次長くしながら可変にし、特徴パラメータをそれぞれにおいて計算する。

## 【0099】

また、第2の標準パターン生成においては、実施の形態6と同様に、第1の単語辞書データベース6022を用いて、データ補間部6032により、全ての標準パターンを生成する。

## 【0100】

実施の形態7の音声認識装置のその他の構成は、実施の形態6の音声認識装置400の構成と同様であるので、その説明は繰り返さない。

## 【0101】

このような実施の形態7の音声認識装置の構成により、音声認識装置300の奏する効果に加え、音韻の平均継続時間長が長い場合は、固定フレーム間隔抽出処理で有効に対処することが可能であり、一方、音韻の平均継続時間長が短い場合は、可変フレーム間隔抽出処理で有効に対処することが可能であるので、処理負荷を低減できる。

## 【0102】

## [実施の形態8]

図10は、実施の形態8の音声認識装置500の構成を説明するための機能ブロック図である。

## 【0103】

図10に示した音声認識装置500の構成においては、入力音声701、入力選択部710、制御信号71、インバータ711、第1の特徴抽出部702、第2の特徴抽出部703、第1の認識処理部7021、第2の認識処理部7031、結果選択部704、第1の単語辞書データベース7022および認識結果705は、それぞれ、実施の形態6の音声認識装置400の入力音声601、入力選択部610、制御信号61、インバータ611、第1の特徴抽出部602、第2の特徴抽出部603、第1の認識処理部6021、第2の認識処理部6031、結果選択部604、第1の単語辞書データベース6022およ

び認識結果 605 に相当する機能を有している。

【0104】

音声認識装置 500 においても、第 2 の特徴抽出部 703 および第 2 の認識処理部 7031 からなる第 2 の処理系でのフレームシフトの時間幅 D2011 が、第 1 の特徴抽出部 702 および第 1 の認識処理部 7021 からなる第 1 の処理系でのフレームシフトの時間幅 D201 よりも長いものであるとする。

【0105】

音声認識装置 500 では、時間幅の最小値は D201 とする。

【0106】

音声認識装置 500 においても、あらかじめ個々の音韻モデルに関して、フレーム長 L から計算した特徴パラメータを用いて、第 1 の標準パターンを作成しておく。この第 1 の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用いて計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅を D201 にして、作成したものである）を、個別の隠れマルコフモデル（HMM）P01 を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数 M の隠れマルコフモデルにより、第 1 の単語辞書データベース 7022 が構成される。

【0107】

第 1 の第 1 の単語辞書データベース 7022 も、図 8 に示したように時刻とパラメータとが関連付けて格納されているものとする。

【0108】

音声認識装置 500 では、第 2 の処理系でフレームシフトの時間幅 D2011 が、第 1 の処理系でのフレームシフトの時間幅 D201 よりも長いだけでなく、長い時間幅 D2011 で変化する際の各時刻点が、短い時間幅 D201 で変化する際の時刻点に相当または対応するように、時間幅 D2011 と時間幅 D201 との関係が定められているものとする。

【0109】

たとえば、時間幅 D2011 での変化に対して、時間幅 D201 の変化が、等比または等差的なものとする場合、第 2 の標準パターンは、実施の形態 6 のような特別な補間操作を必要とせずに、第 1 の標準パターンから得ることができる。

【0110】

実施の形態 8 の音声認識装置のその他の構成および動作は、実施の形態 6 の音声認識装置 400 の構成と同様であるので、その説明は繰り返さない。

【0111】

このような実施の形態 8 の音声認識装置の構成により、音声認識装置 400 の奏する効果に加え、一層、処理負荷を低減できる。

【0112】

今回開示された実施の形態はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は上記した説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

【図面の簡単な説明】

【0113】

【図 1】 音声認識装置 10 の構成を説明するための機能ブロック図である。

【図 2】 図 1 に示した特徴検出部 102 におけるフレームシフトを説明するための概念図である。

【図 3】 音声認識装置 100 の構成を説明するための機能ブロック図である。

【図 4】 音声認識装置 100 の特徴パラメータ計算部 3021 のフレームシフト動作を説明するための概念図である。

【図 5】 実施の形態 2 の音声認識装置 200 の構成を説明するための機能ブロック図



である。

【図 6】実施の形態 4 の音声認識装置 3 0 0 の構成を説明するための機能ブロック図である。

【図 7】実施の形態 6 の音声認識装置 4 0 0 の構成を説明するための機能ブロック図である。

【図 8】標準パターンが、第 1 の単語辞書データベース 6 0 2 2 に格納される状態を説明するための概念図である。

【図 9】データ補間部 6 0 3 2 の処理を説明するための概念図である。

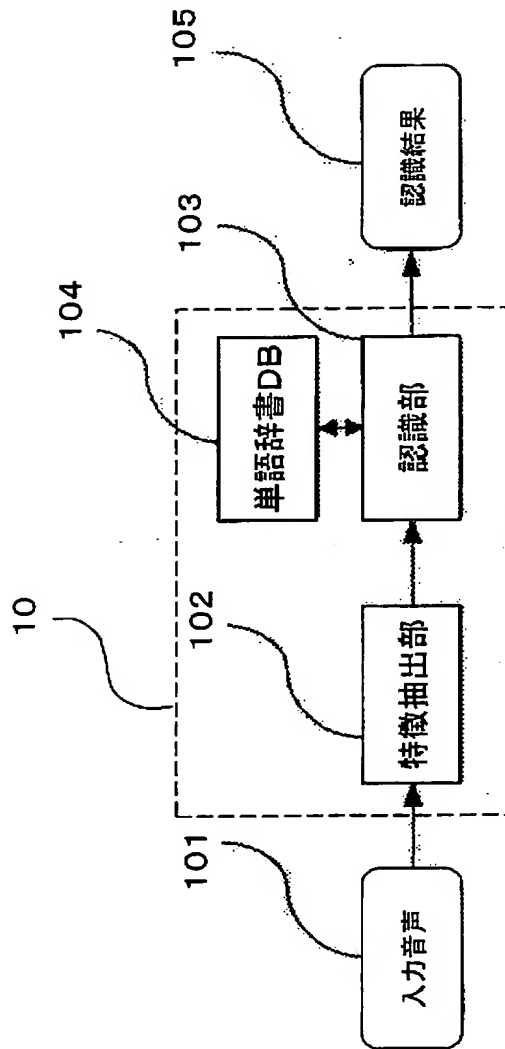
【図 1 0】実施の形態 8 の音声認識装置 5 0 0 の構成を説明するための機能ブロック図である。

【符号の説明】

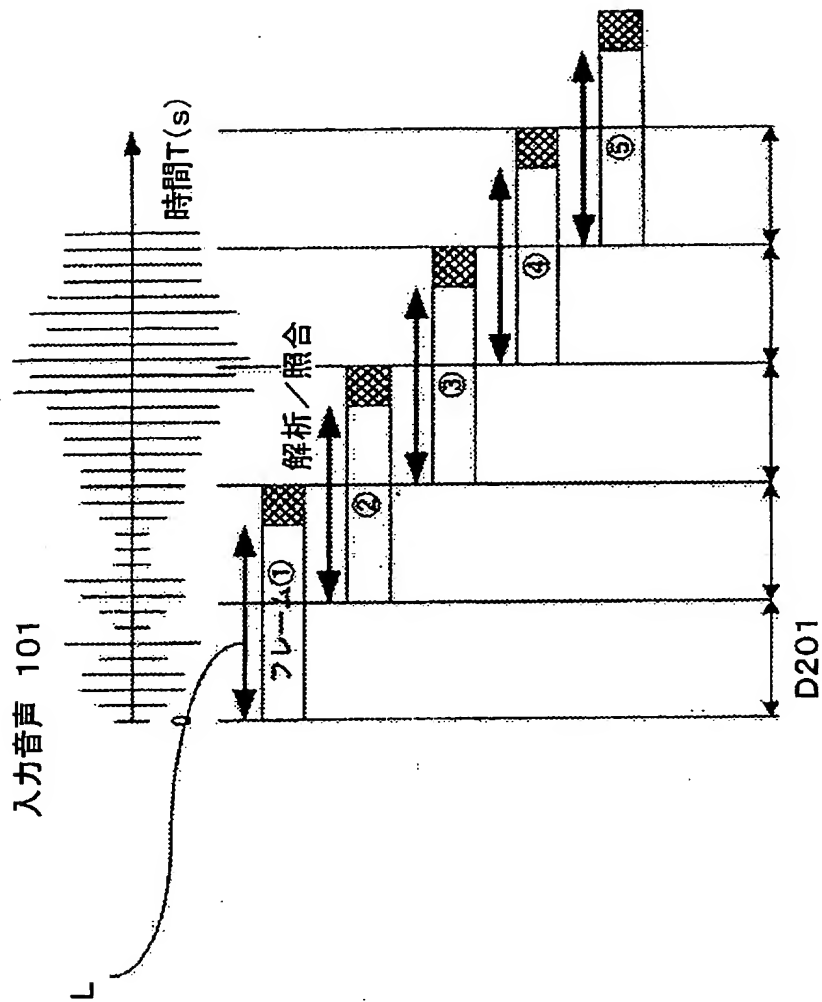
【0 1 1 4】

5 1, 6 1, 7 1 制御信号、1 0 1, 3 0 1, 4 0 1, 5 0 1, 6 0 1, 7 0 1 入力音声、5 1 0, 6 1 0, 7 1 0 入力選択部、3 0 2 特徴抽出部、5 1 1, 6 1 1, 7 1 1 インバータ、4 0 2, 5 0 2, 6 0 2, 7 0 2 第 1 の特徴抽出部、4 0 3, 5 0 3, 6 0 3, 7 0 3 第 2 の特徴抽出部、4 0 2 1, 5 0 2 1, 6 0 2 1, 7 0 2 1 第 1 の認識処理部、4 0 3 1, 5 0 3 1, 6 0 3 1, 7 0 3 1 第 2 の認識処理部、4 0 4, 5 0 4, 6 0 4, 7 0 4 結果選択部、4 0 2 2, 5 0 2 2, 6 0 2 2, 7 0 2 2 第 1 の単語辞書データベース、4 0 3 2, 5 0 3 2 第 2 の単語辞書データベース、6 0 3 2 データ補間部、1 0 5, 3 0 5, 4 0 5, 5 0 5, 6 0 5, 7 0 5 認識結果、1 0、1 0 0, 2 0 0, 3 0 0, 4 0 0, 5 0 0 音声認識装置。

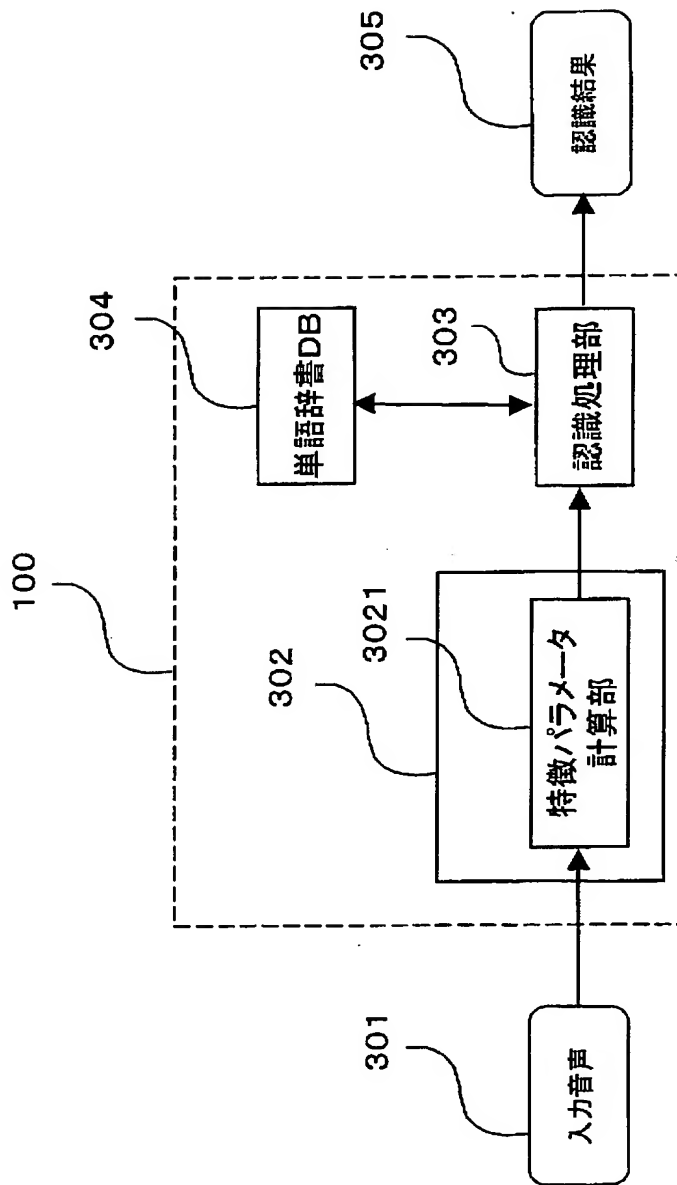
【書類名】 図面  
【図 1】



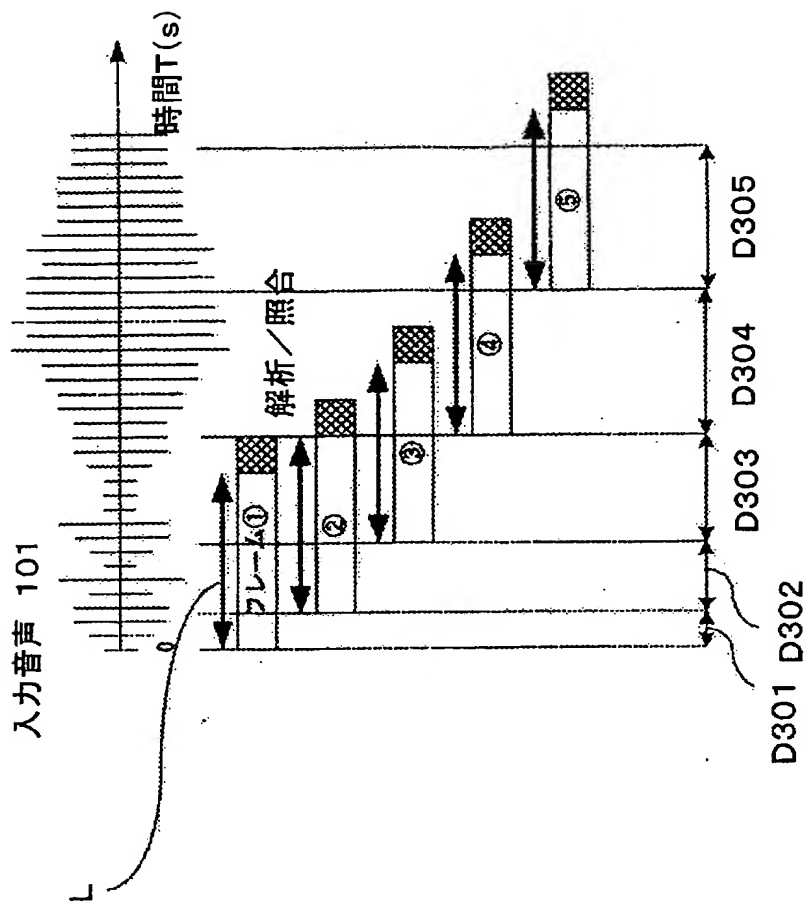
【図 2】



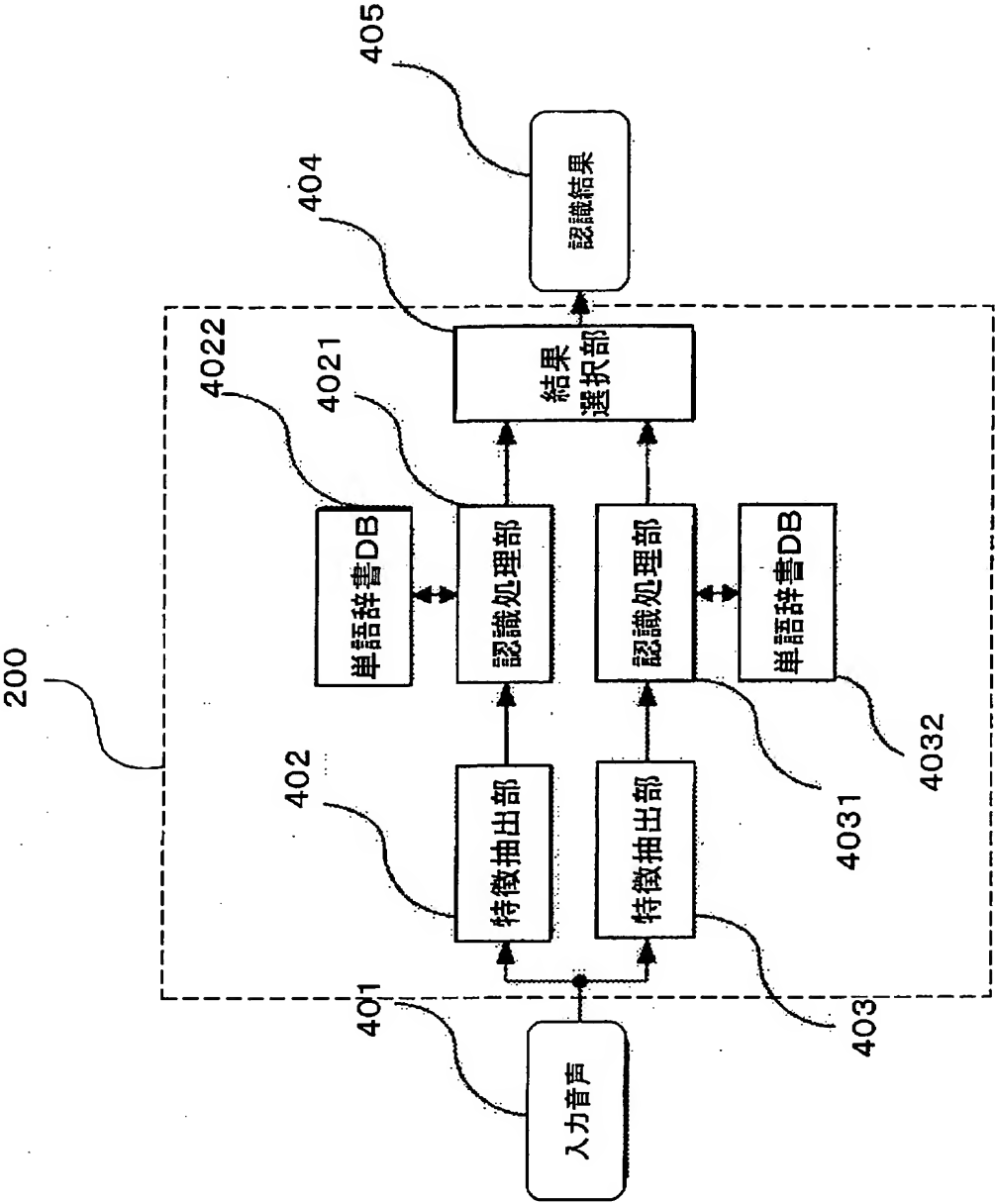
【図 3】



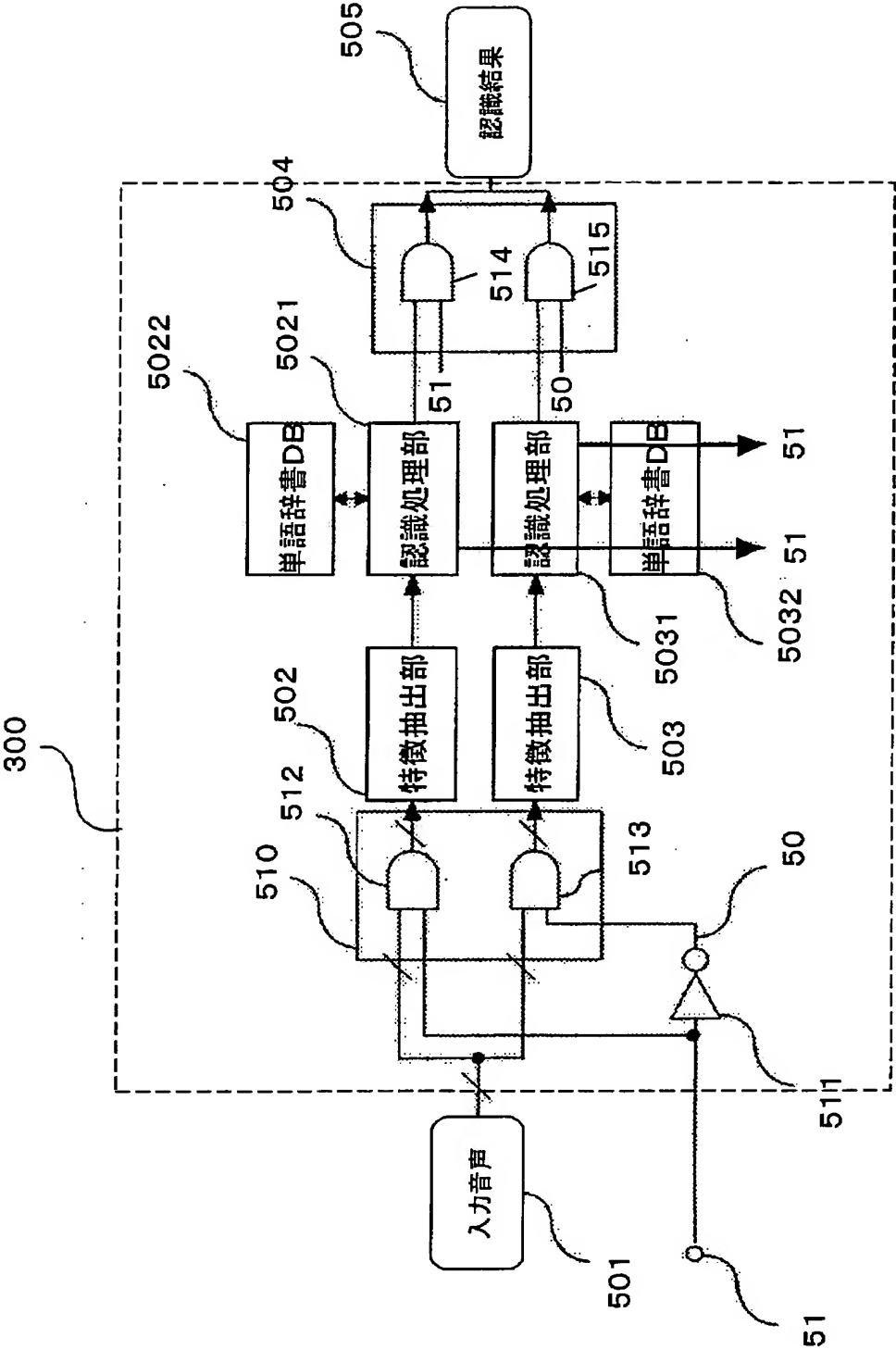
【図 4】



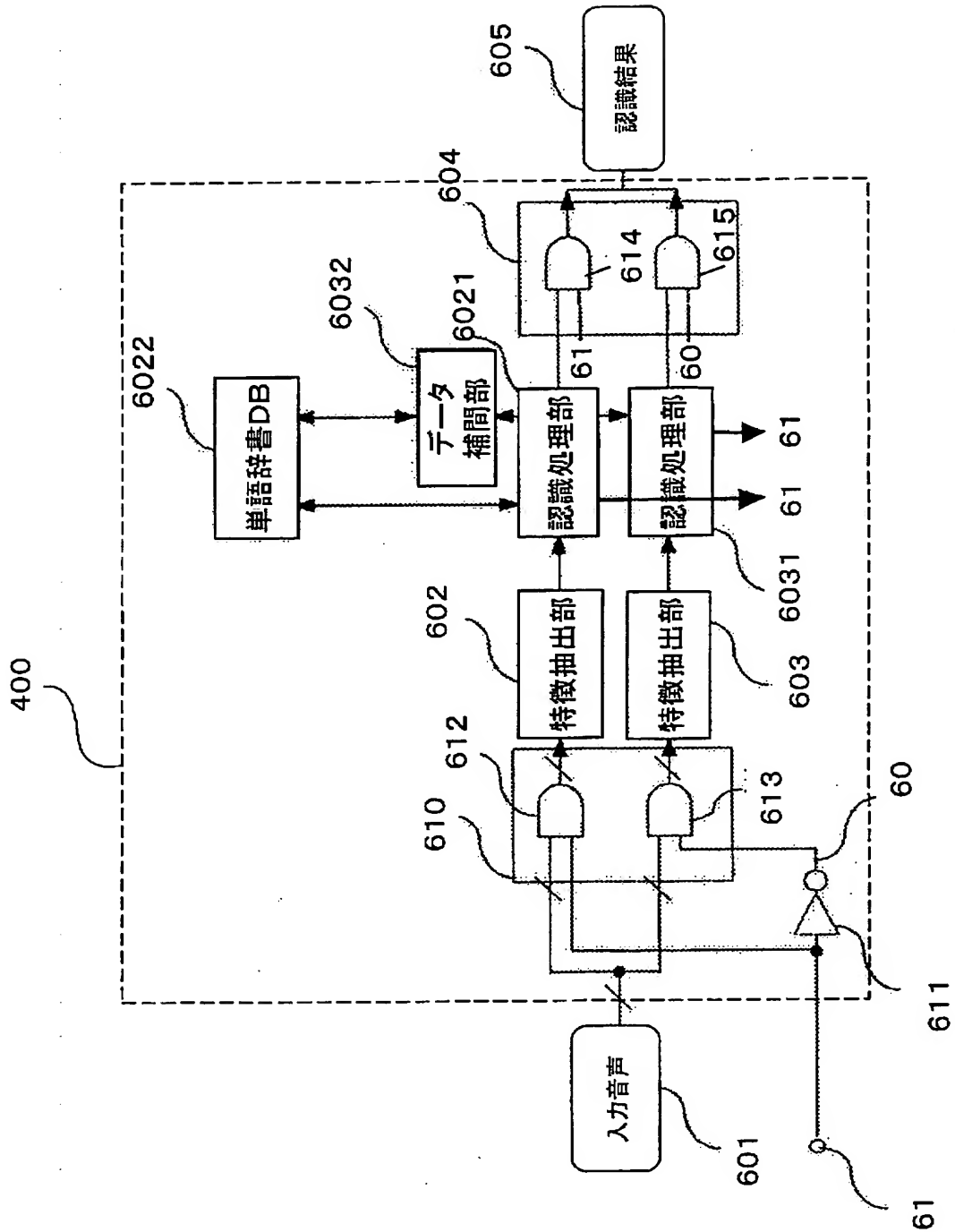
【図 5】



【図 6】

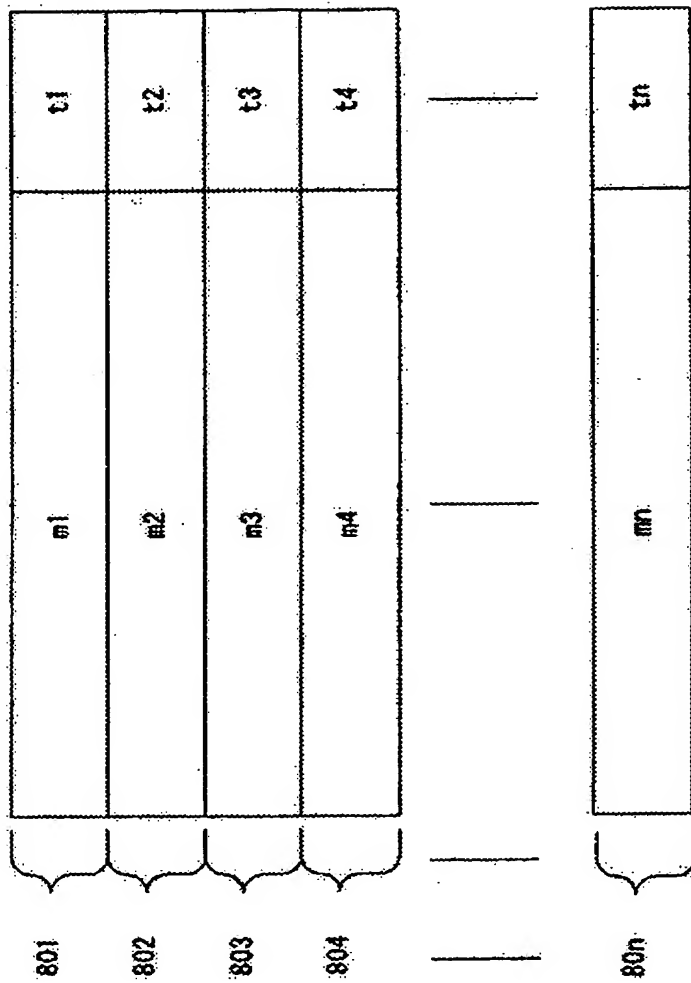


【図 7】



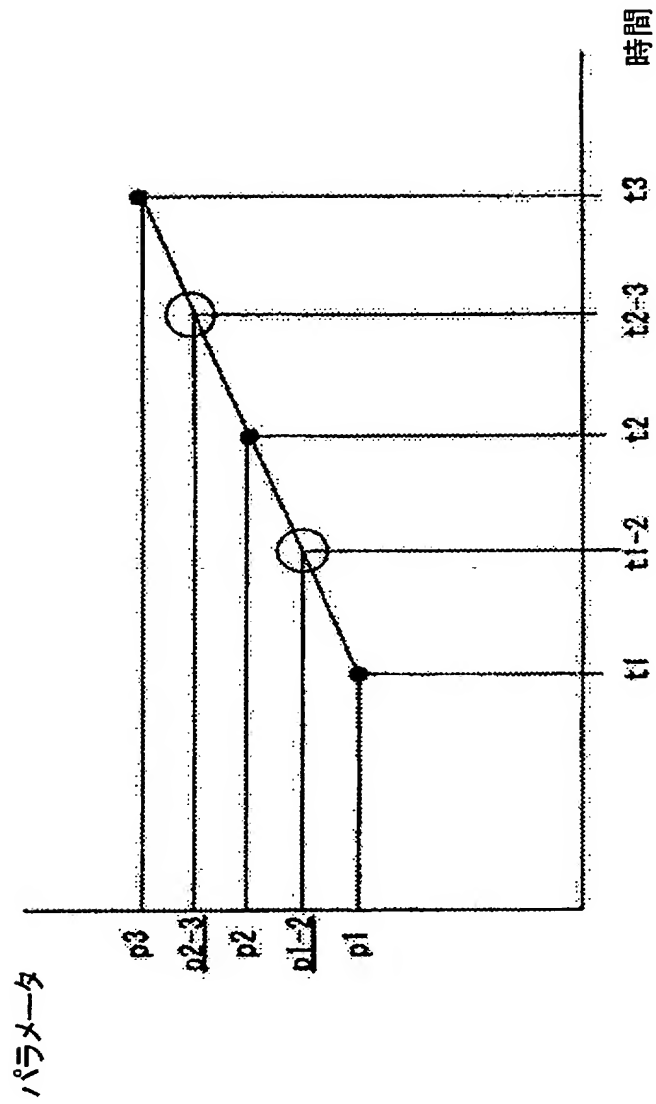


【図 8】

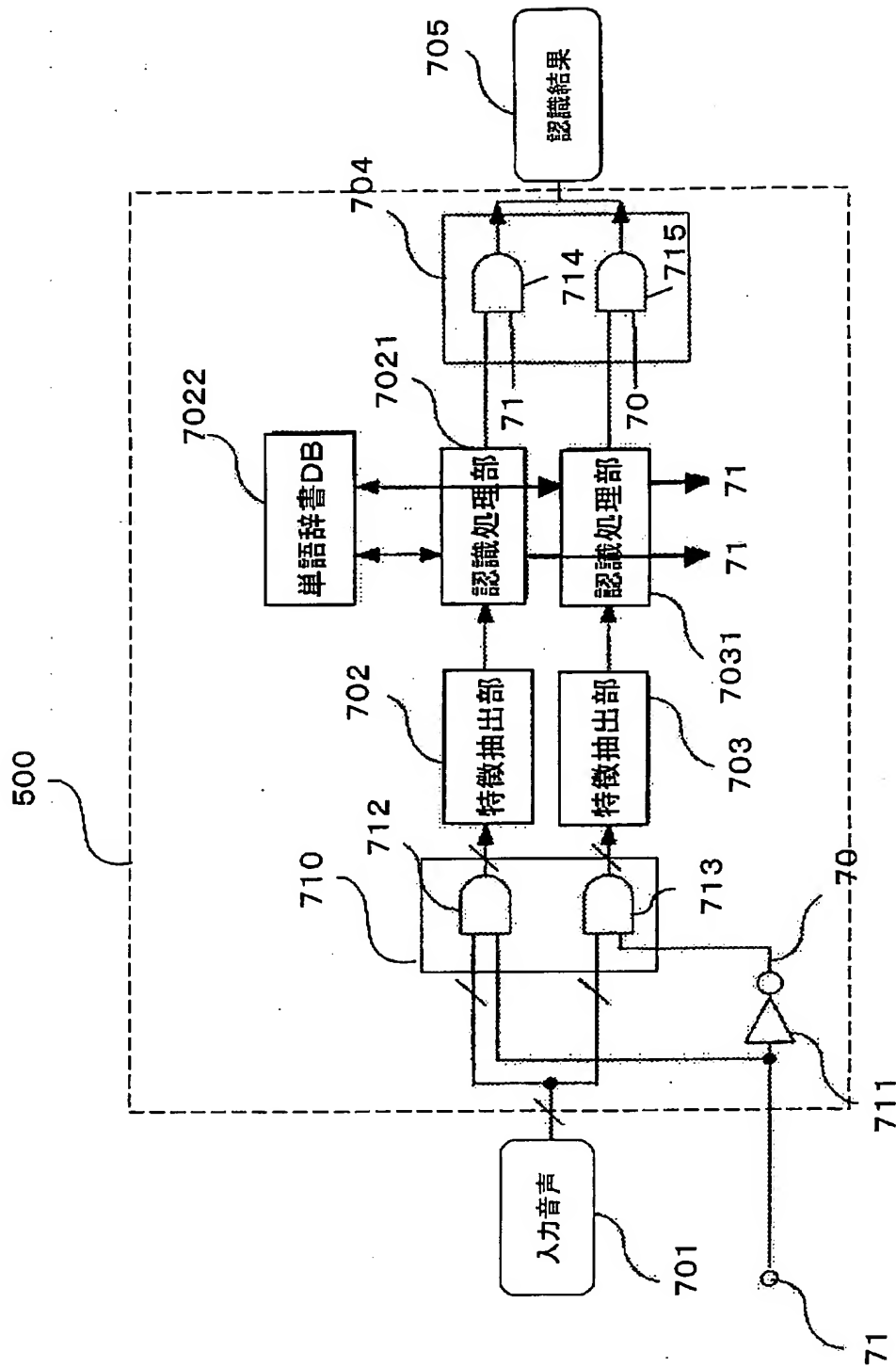


800

【図 9】



【図10】



【書類名】 要約書

【要約】

【課題】 各音韻の認識率を向上することが可能な特徴パラメータの計算方法を用いた音声認識装置を提供する。

【解決手段】 音声認識装置 1 0 において、特徴抽出部 3 0 2 は、入力音声信号 3 0 1 に対して、各々が所定時間長を有する時間窓に対応する複数のフレームを、順次増加する時間幅でスライドさせることにより、特徴パラメータの抽出を行う。単語辞書データベース 3 0 4 は、入力音声の音韻パターンにそれぞれ対応づけて標準パターンデータを格納する。認識処理部 3 0 3 は、特徴抽出部 3 0 2 で抽出された特徴パラメータと標準パターンデータを照合することで、対応する音韻を認識し、認識結果を出力する。

【選択図】 図 3

特願 2 0 0 3 - 2 7 7 6 6 1

出 願 人 履 歴 情 報

識別番号

[ 5 0 3 1 2 1 1 0 3 ]

1. 変更年月日

2 0 0 3 年 4 月 1 日

[変更理由]

新規登録

住 所

東京都千代田区丸の内二丁目 4 番 1 号

氏 名

株式会社ルネサステクノロジ